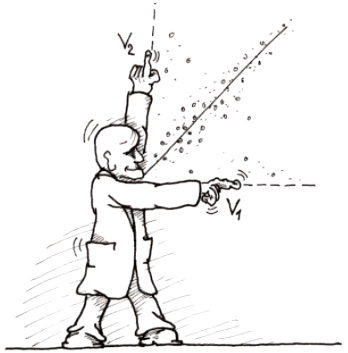
Statistics 203

Thomas Rieg

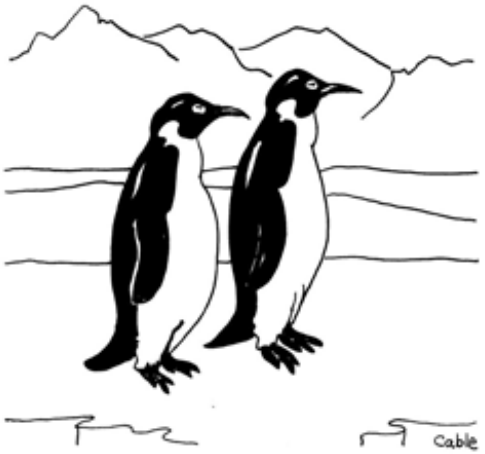
Clinical Investigation & Research Department

Naval Medical Center Portsmouth



Correlation and Regression

- Correlation
- Regression
- Logistic Regression

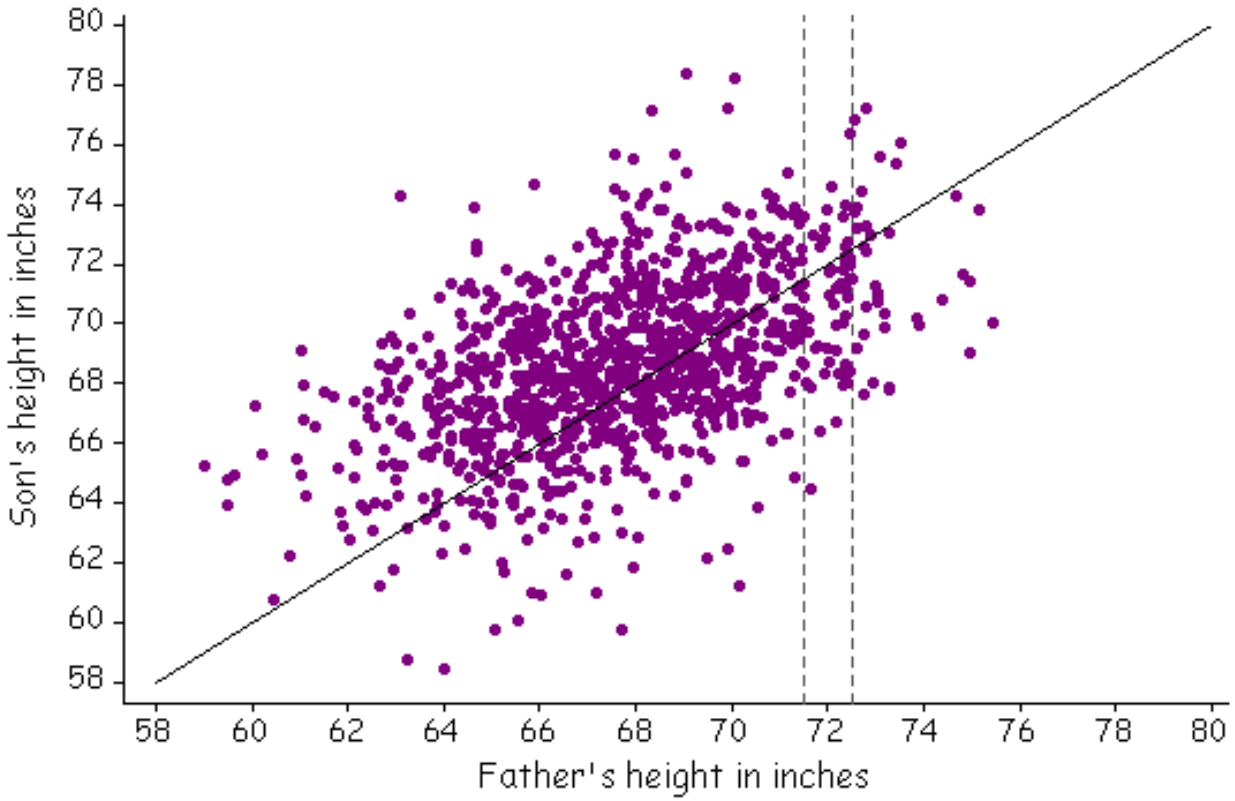


"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"

History

- Karl Pearson (1857-1936) considered the data corresponding to the heights of 1,078 fathers and their son's at maturity
- A list of these data is difficult to understand, but the relationship between the two variables can be visualized using a scatter diagram, where each pair father-son is represented as a point in a plane
- The x-coordinate corresponds to the father's height and the y-coordinate to the son's
- The taller the father the taller the son
- This corresponds to a positive association
- He considered the height of the father as an independent variable and the height of the son as a dependent variable

Pearson's Data



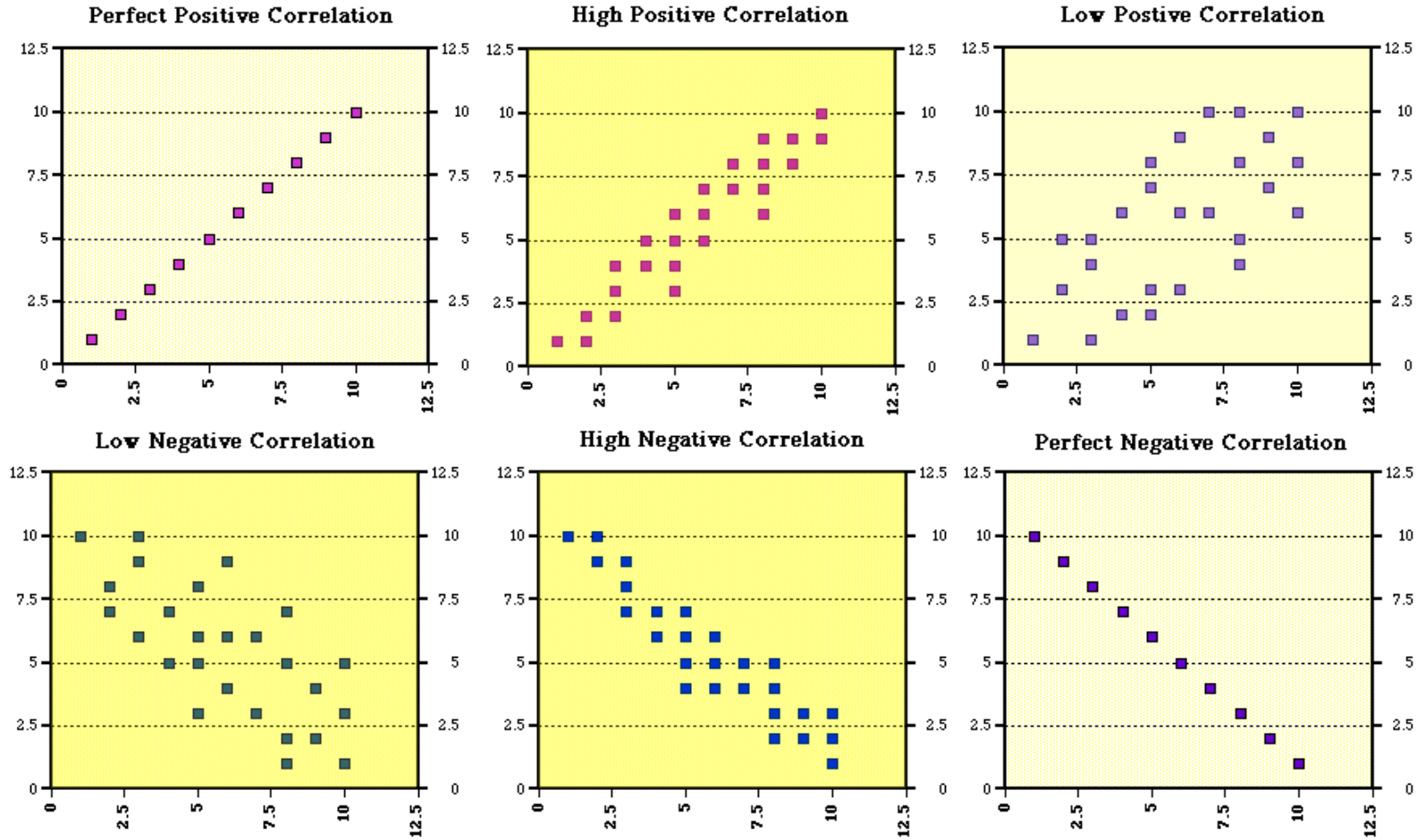
Galton's data

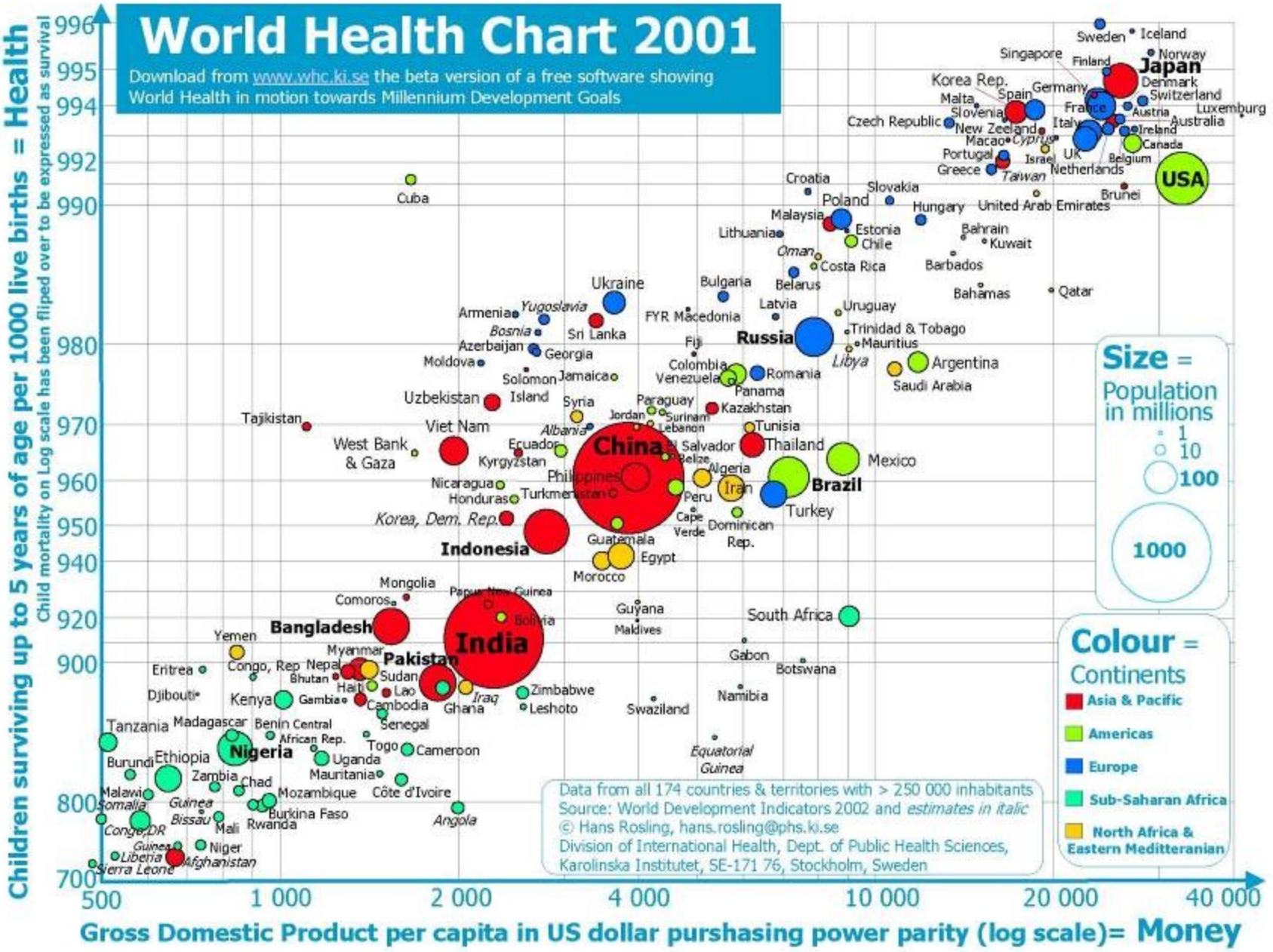
- What do the data show?
- The taller the father, the taller the son
  - Tall father's son is taller than short father's son
- But tall father's son is not as tall as father
  - Short father's son is not as short as father

Correlation

- The correlation gives a measure of the linear association between two variables
- To what degree are two things related
- It is a coefficient that does not depend on the units that are used to measure the data
- And is bounded between -1 and 1

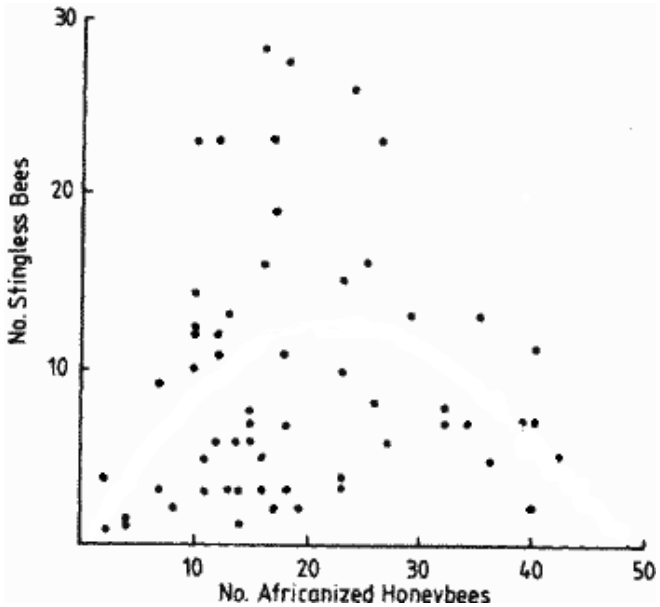
Scatterplots

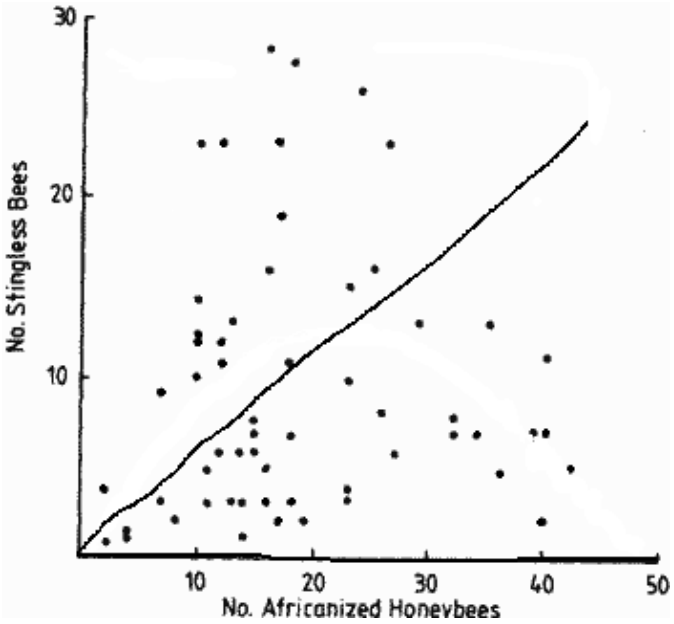




www.gapminder.org

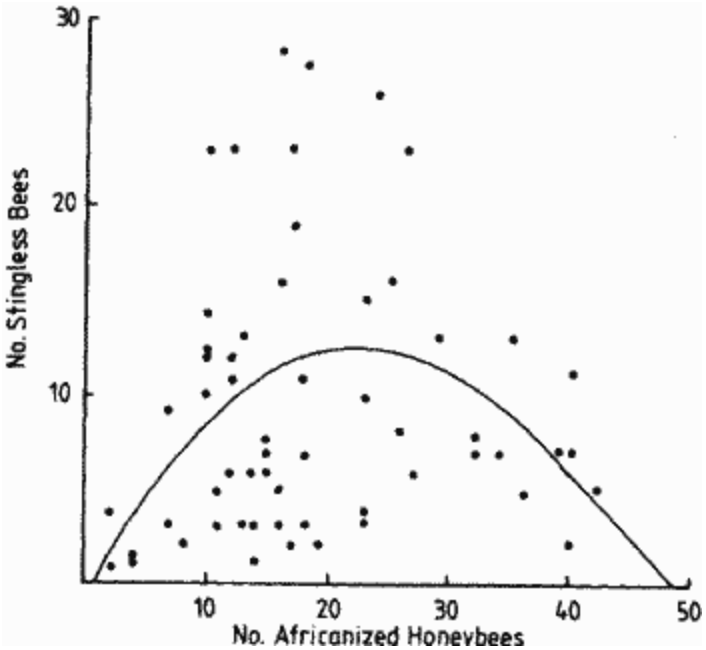
Curve Fitting

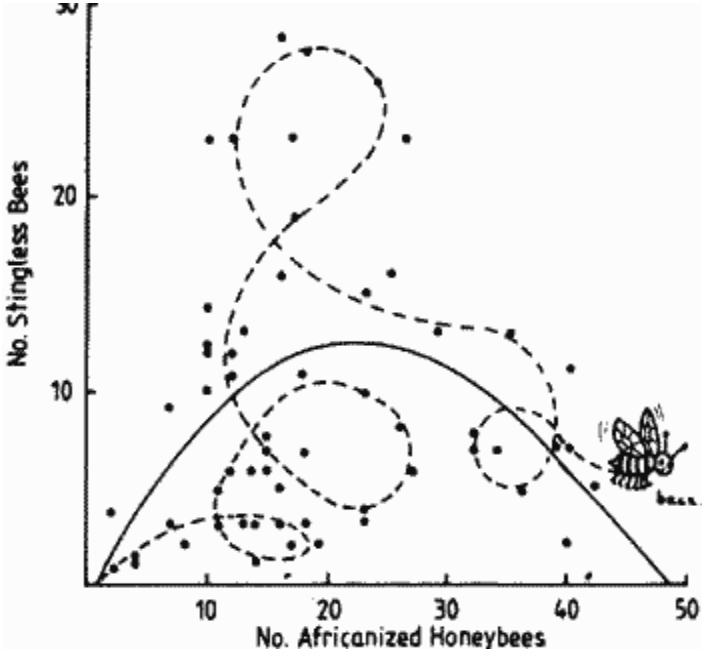




- Roubik (Science 1978: 201;1030)

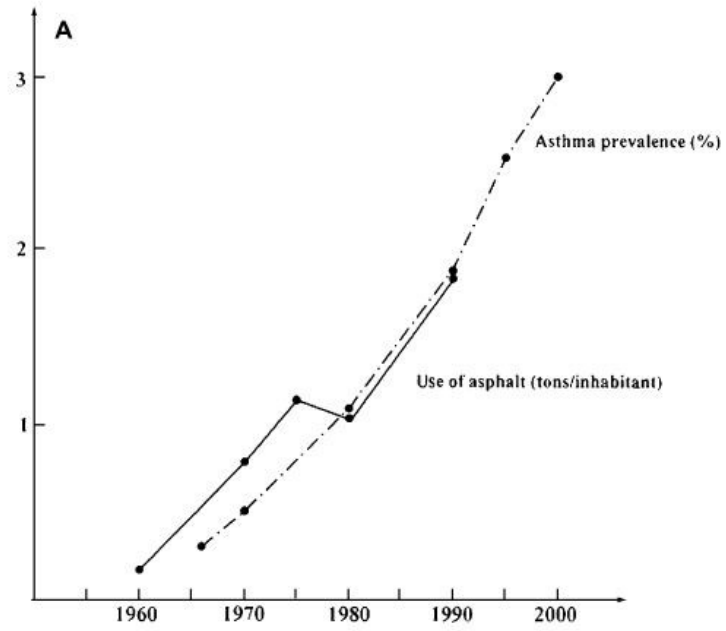
More Curve Fitting



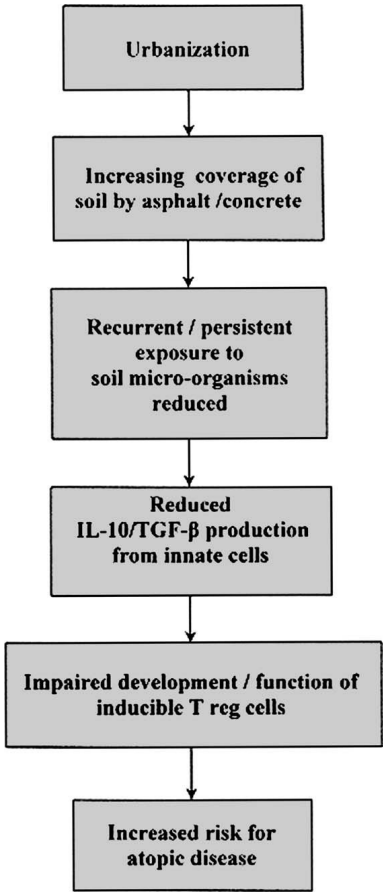


- Roubik (Science 1978: 201;1030)

Correlational Approach



Leena von Hertzen, & Tari Haahtela. (2006). Disconnection of man and the soil: Reason for the asthma and atopy epidemic? Journal of Allergy and Clinical Immunoloty, 117(2), 334-344.


Urbanization
Increasing coverage of soil by asphalt /concrete
Recurrent / persistent exposure to soil micro-organisms reduced
Reduced IL-10/TGF-β production from innate cells
Impaired development / function of inducible T reg cells
Increased risk for atopic disease

Causation

- The more bars a city has the more churches it has as well
  - Religion causes drinking?
- Students with tutors have lower test scores
  - Tutoring lowers test scores?
- Near Perfect Correlation:

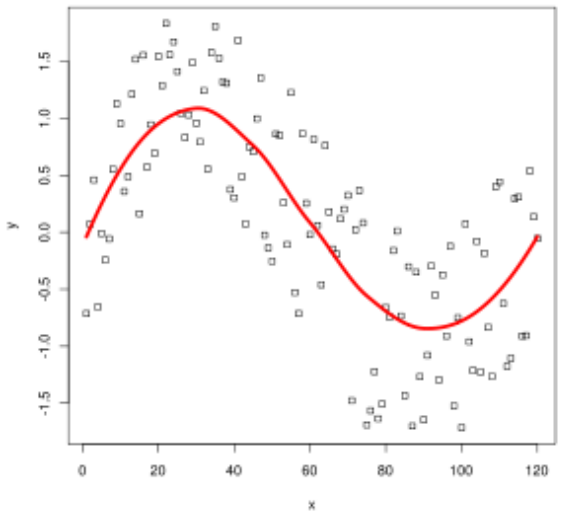
  Kissing and Pregnancy

Types of Correlations

Point-biserial r	One dichotomous variable (yes/no; male/female) and one interval or ratio variable
Biserial r	One variable forced into a dichotomy (grade distribution dichotomized to “pass” and “fail”) and one interval or ratio variable
Phi coefficient	Both variables are dichotomous on a nominal scale (male/female *vs.* high school graduate/dropout)
Tetrachoric r	Both variables are dichotomous with underlying normal distributions (pass/fail on a test *vs.* tall/short in height)
Correlation ratio	There is a curvilinear rather than linear relationship between the variables (also called the eta coefficient)
Partial correlation	The relationship between two variables is influenced by a third variable (*e.g.,* mental age and height, which is influenced by chronological age)
Multiple R	The maximum correlation between a dependent variable and a combination of independent variables (a college freshman’s GPA as predicted by his high school grades in Math, chemistry, history, and English)

Usefullness of Correlation

- Correlation is useful only when measuring the degree of linear association between two variables.
- That is, how much the values from two variables cluster around a straight line



- The variables in this plot have an obvious nonlinear association
- Nevertheless the correlation between them is 0.3
- This is because the points are clustered around a sinus curve and not a straight line

Linear Regression

- **Correlation**
  - measures the *degree of association* between variables
- **Linear Regression** is a development of the *Pearson Product Moment correlation*
- **Bivariate** (Two Variable) **Regression** *plus* **Multiple Regression**: two or more variables
- Both **Correlation** and **Regression** Analysis will tell you if there is a **significant** relationship between variables and both provide an index of the **strength** of that relationship

Introduction to Regression Analysis

- Regression analysis is the most often applied technique of statistical analysis and modeling
- In general, it is used to model a response variable (Y) as a function of one or more driver variables (X_1, X_2, ..., X_p)
- The functional form used is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon$$

Introduction to Regression Analysis

- If there is only one driver variable, X, then we usually speak of “simple” linear regression analysis
- When the model involves
  - (a) multiple driver variables,
  - (b) a driver variable in multiple forms, or
  - (c) a mixture of these,

  Then we speak of “multiple linear regression analysis”

- The “linear” portion of the terminology refers to the response variable being expressed as a “linear combination” of the driver variables.

Introduction to Regression Analysis (RA)

- Regression Analysis is used to estimate a function $f(\ )$ that describes the relationship between a continuous dependent variable and one or more independent variables

$$Y = f(X_1, X_2, X_3, \ldots, X_n) + \varepsilon$$

Note:

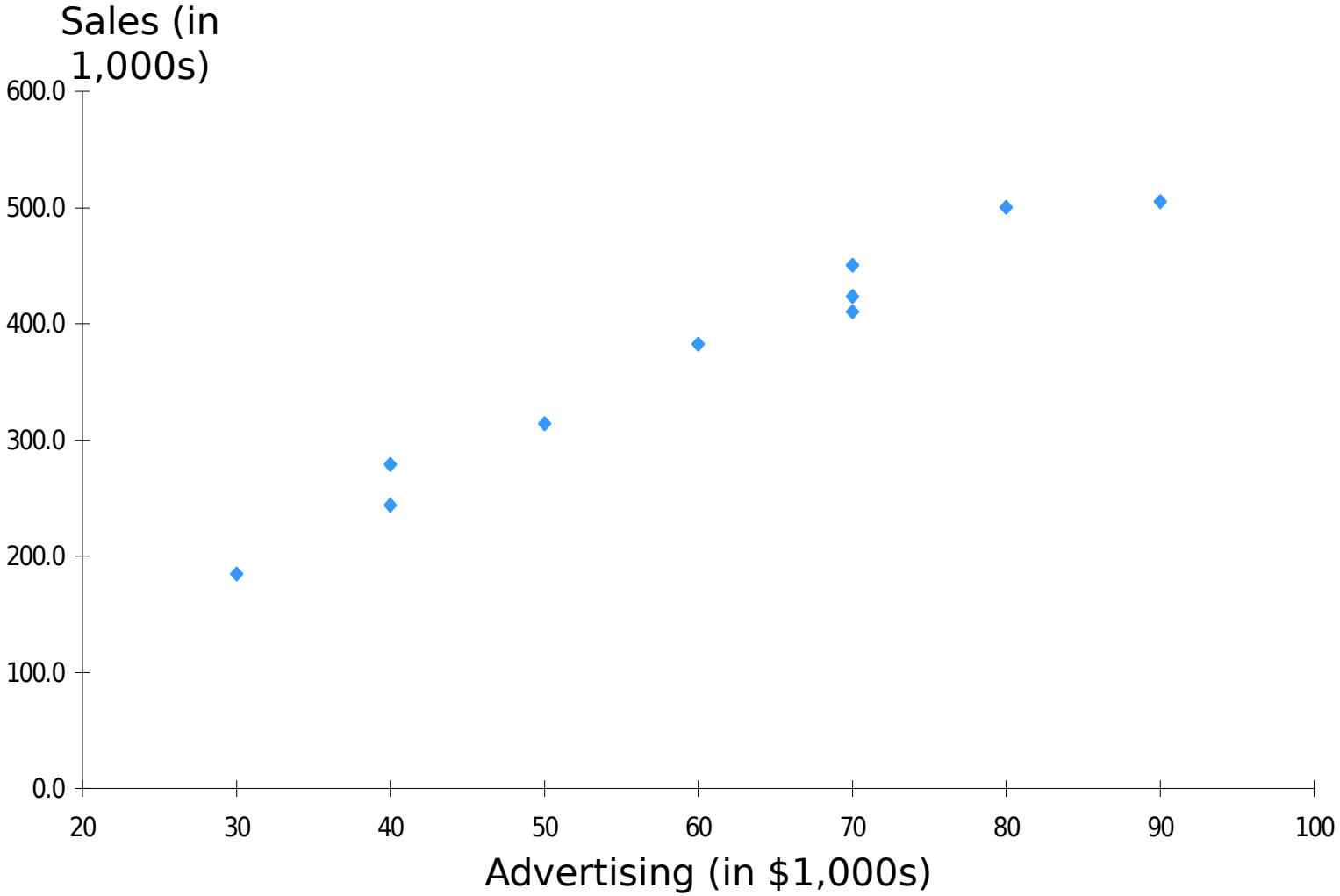
- $f(\ )$ describes *systematic* variation in the relationship
- ε represents the *unsystematic* variation (or random error) in the relationship

An Example

- Consider the relationship between advertising (X_1) and sales (Y) for a company
- There probably *is* a relationship...

  ...as advertising increases, sales should increase
- But how would we measure and quantify this relationship?

A Scatter Plot of the Data



A Simple Linear Regression Model

- The scatter plot shows a linear relation between advertising and sales
- So the following regression model is suggested by the data,

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \varepsilon_i$$

This refers to the true relationship between the entire population of advertising and sales values.

- The estimated regression function (based on our sample) will be represented as,

$$\hat{Y}_i = b_0 + b_1 X_{1_i}$$

$\hat{Y}_i$ is the estimated (or fitted) value of Y at a given level of X

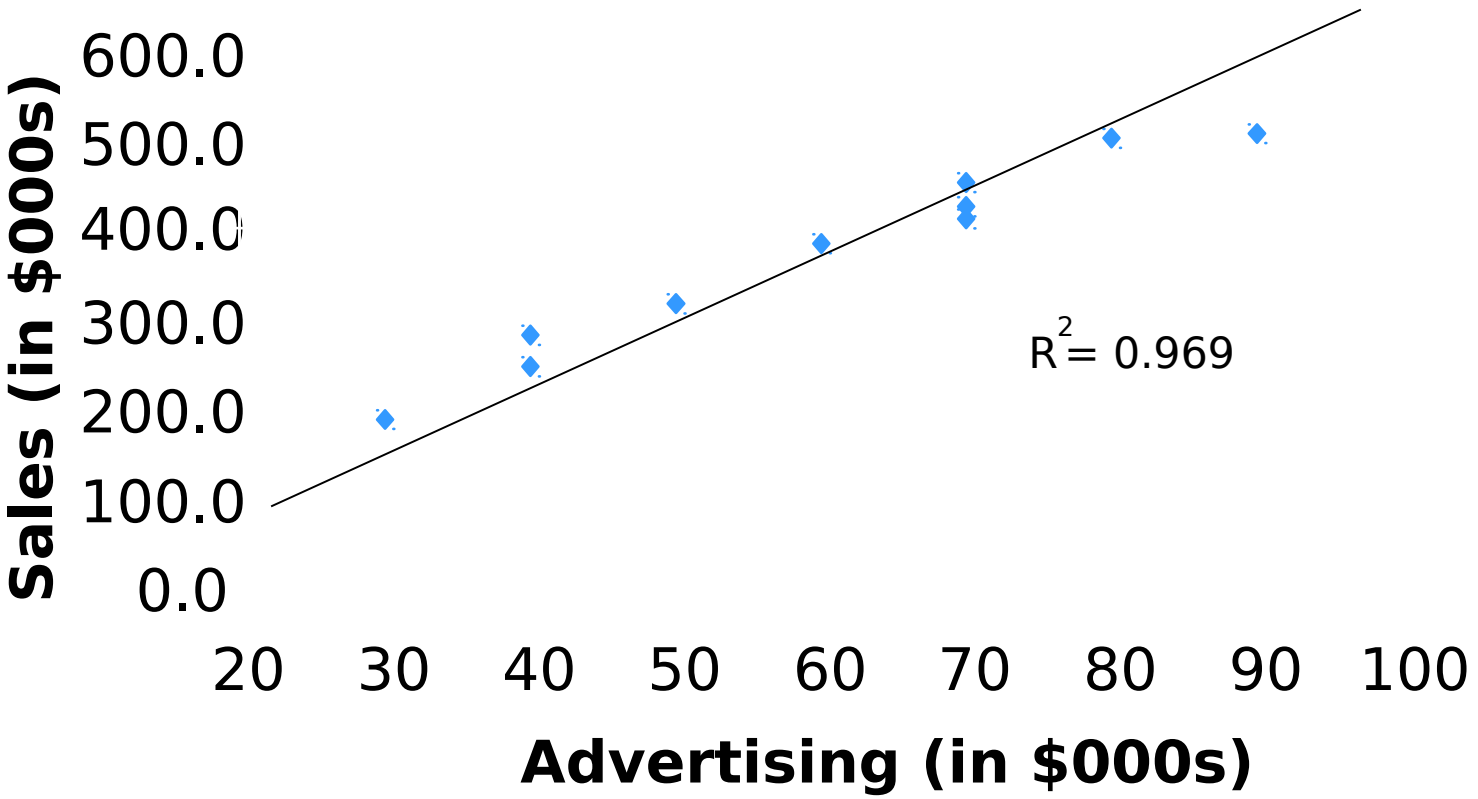
Determining the Best Fit

- Numerical values must be assigned to b_0 and b_1
- The method of “least squares” selects the values that minimize:

$$\mathrm{ESS} = \sum_{i=1}^{n} (\mathrm{Y}_i - \hat{\mathrm{Y}}_i)^2 = \sum_{i=1}^{n} (\mathrm{Y}_i - (b_0 + b_1 \mathrm{X}_{1_i}))^2$$

- If ESS = 0 our estimated function fits the data perfectly

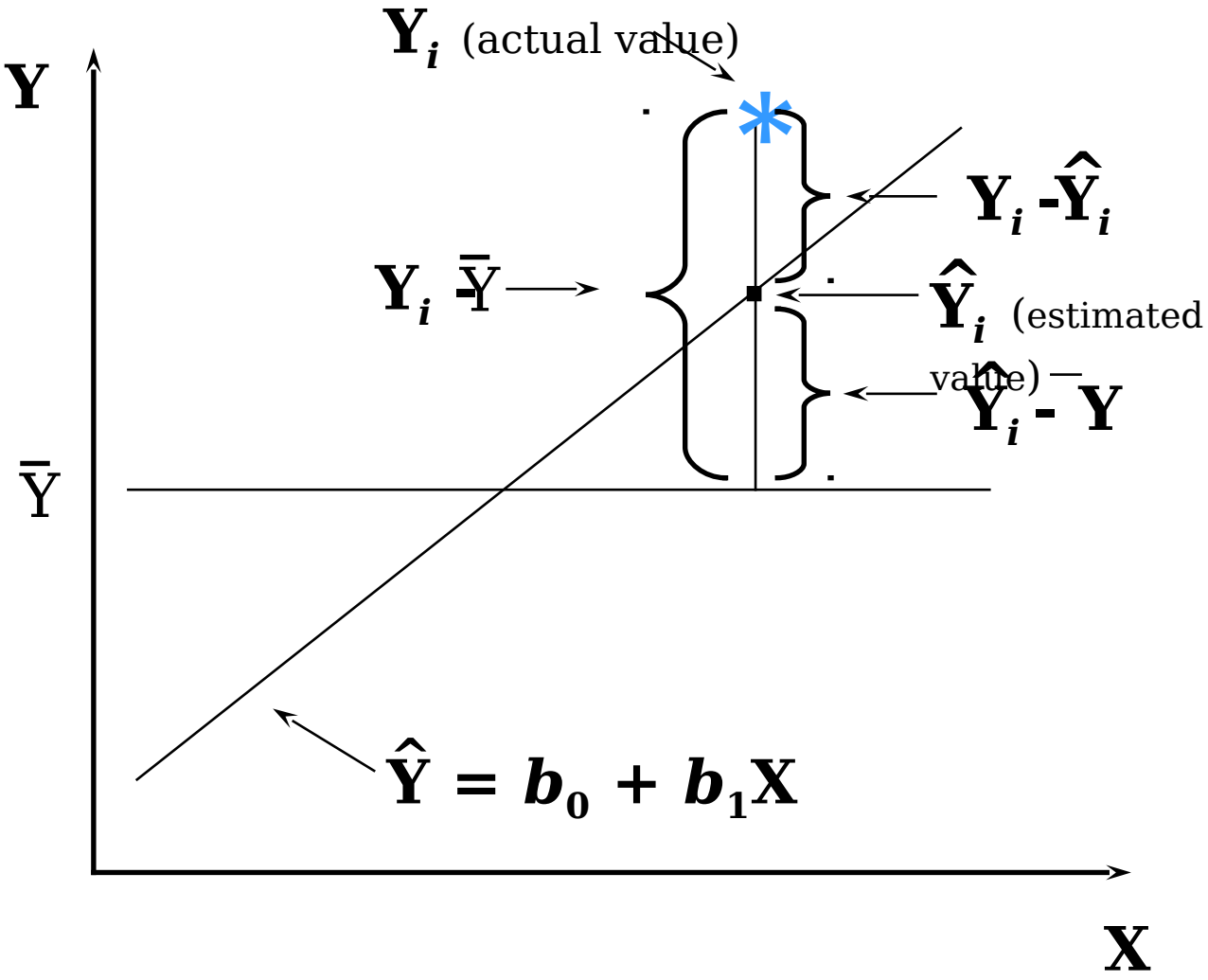
Evaluating the "Fit"



The R^2 Statistic

- The R^2 statistic indicates how well an estimated regression function fits the data
- $0 <= R^2 <= 1$
- It measures the proportion of the total variation in Y around its mean that is accounted for by the estimated regression equation
- To understand this better, consider the following graph . . .

Error Decomposition



Partition of the Total Sum of Squares

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

or,

$$\mathbf{TSS = ESS + RSS}$$

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

Making Predictions

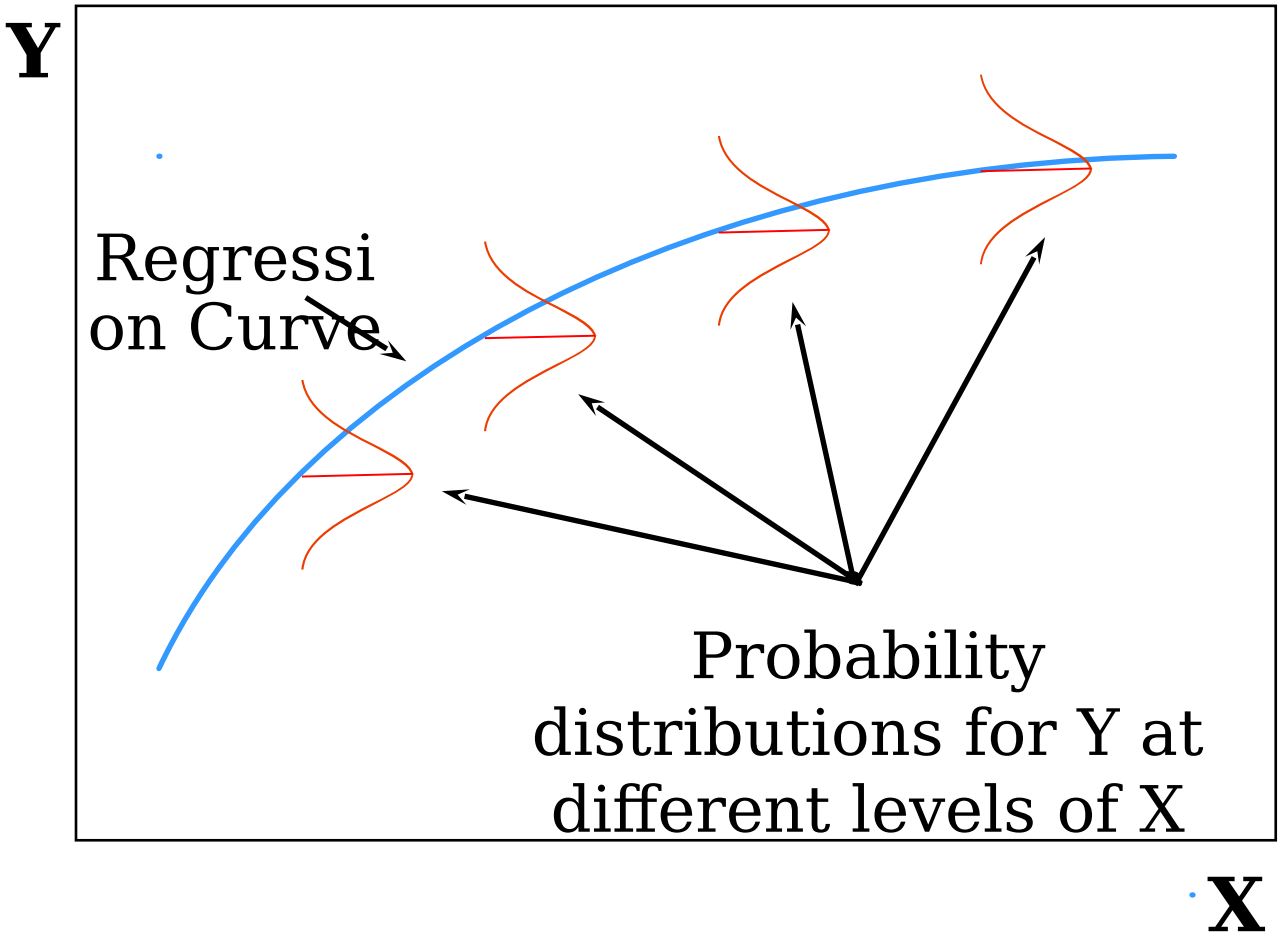
- Suppose we want to estimate the average levels of sales expected if $65K is spent on advertising

$$\hat{Y}_i = 36.342 + 5.550X_{1_i}$$

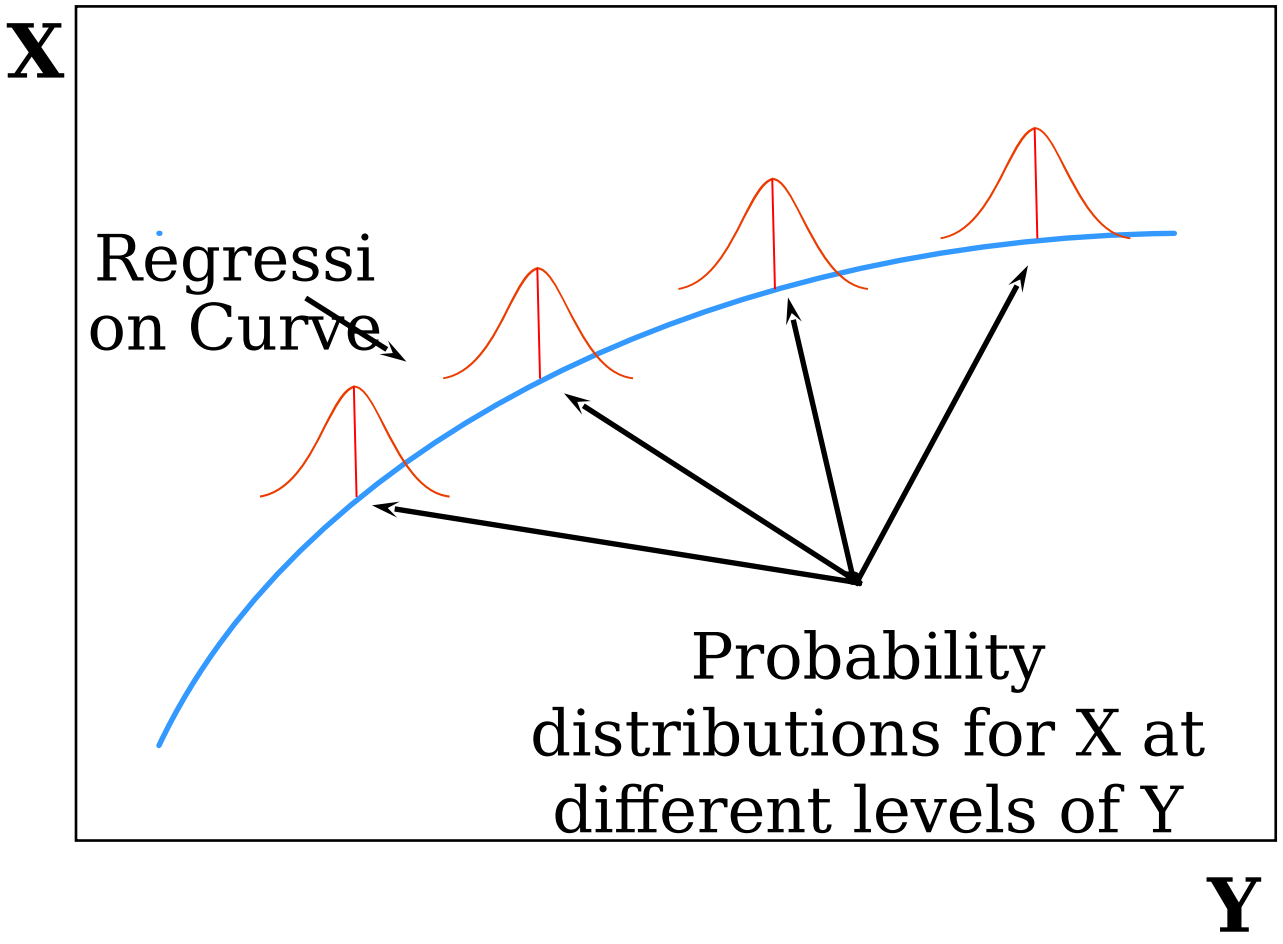
  - Estimated Sales = 36.342 + 5.550 * 65

    = 397.092
  - So when $65,000 is spent on advertising, we expect the average sales level to be $397,092.

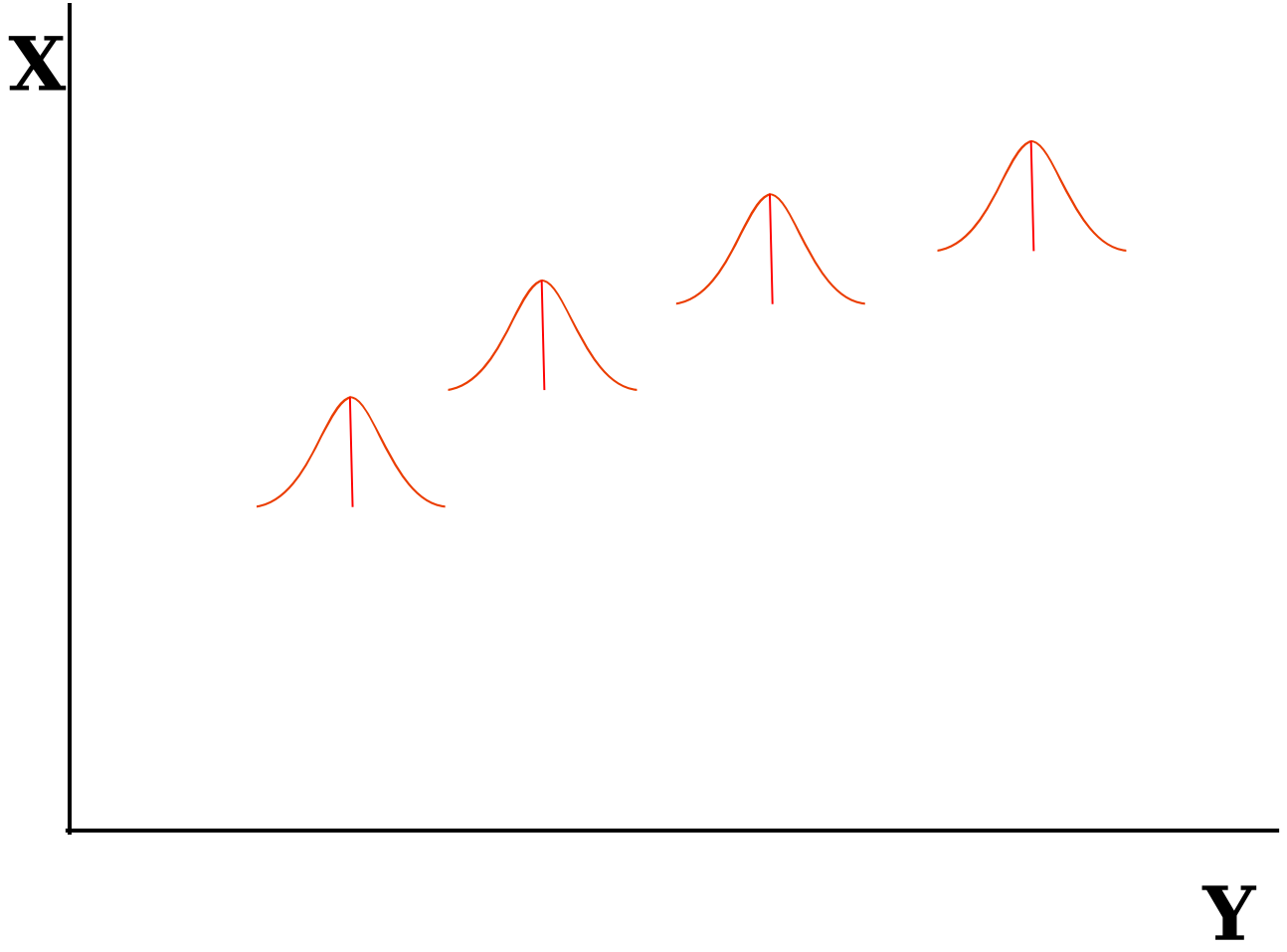
Nature of Statistical Relationship



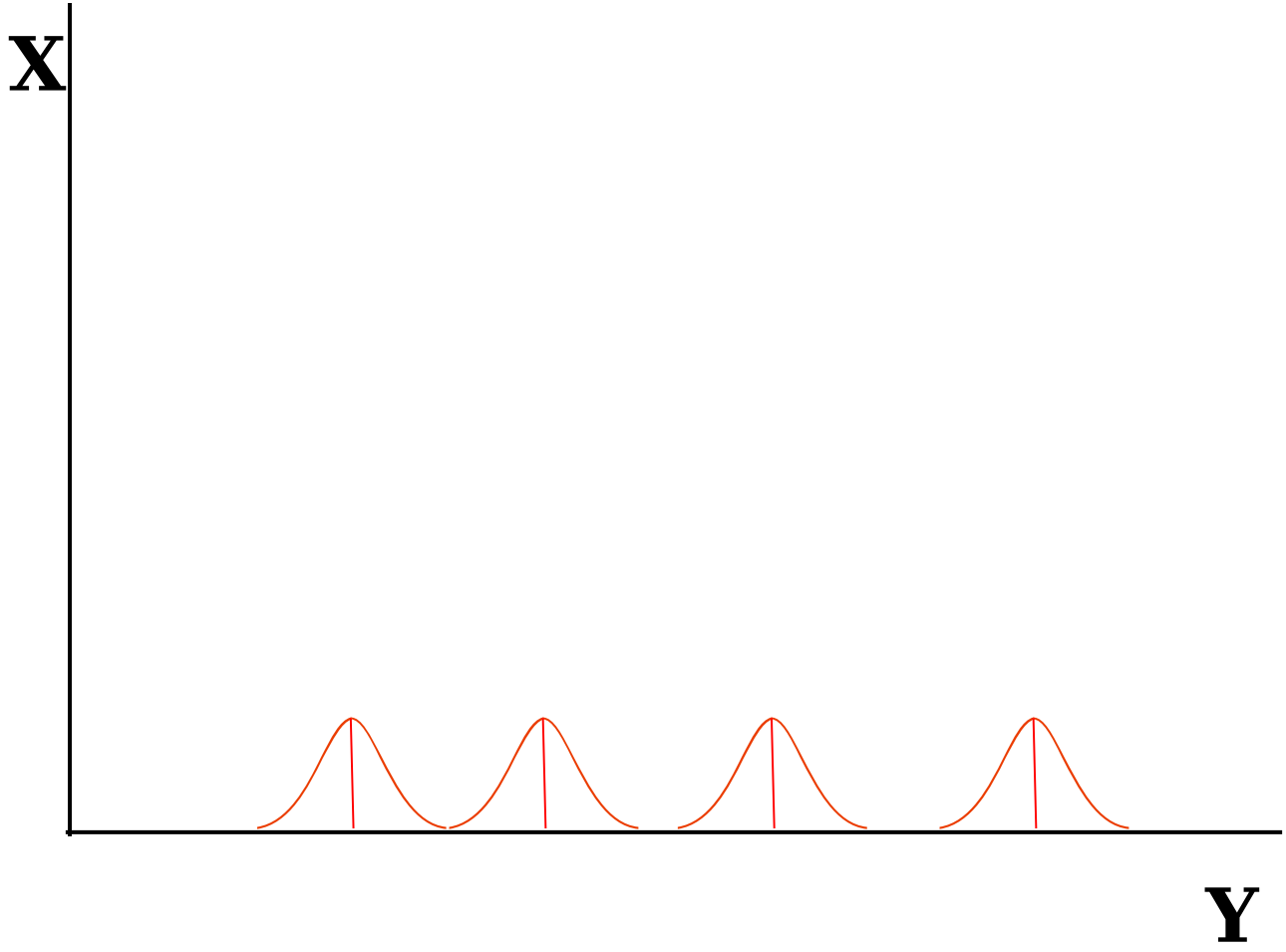
Nature of Statistical Relationship



Nature of Statistical Relationship



Nature of Statistical Relationship

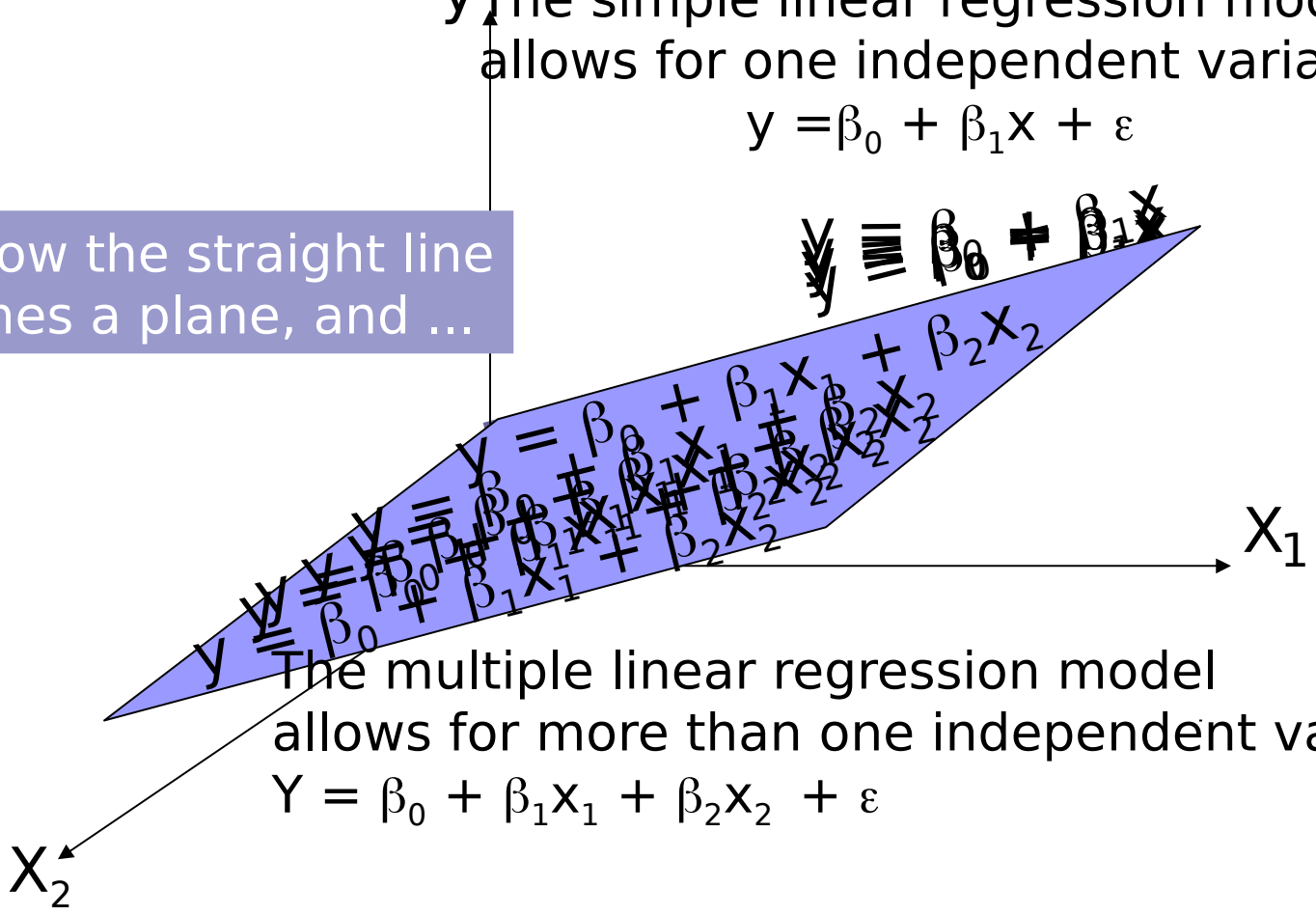


Multiple Regression for k = 2

The simple linear regression model allows for one independent variable, “x

$$y = \beta_0 + \beta_1 x + \varepsilon$$

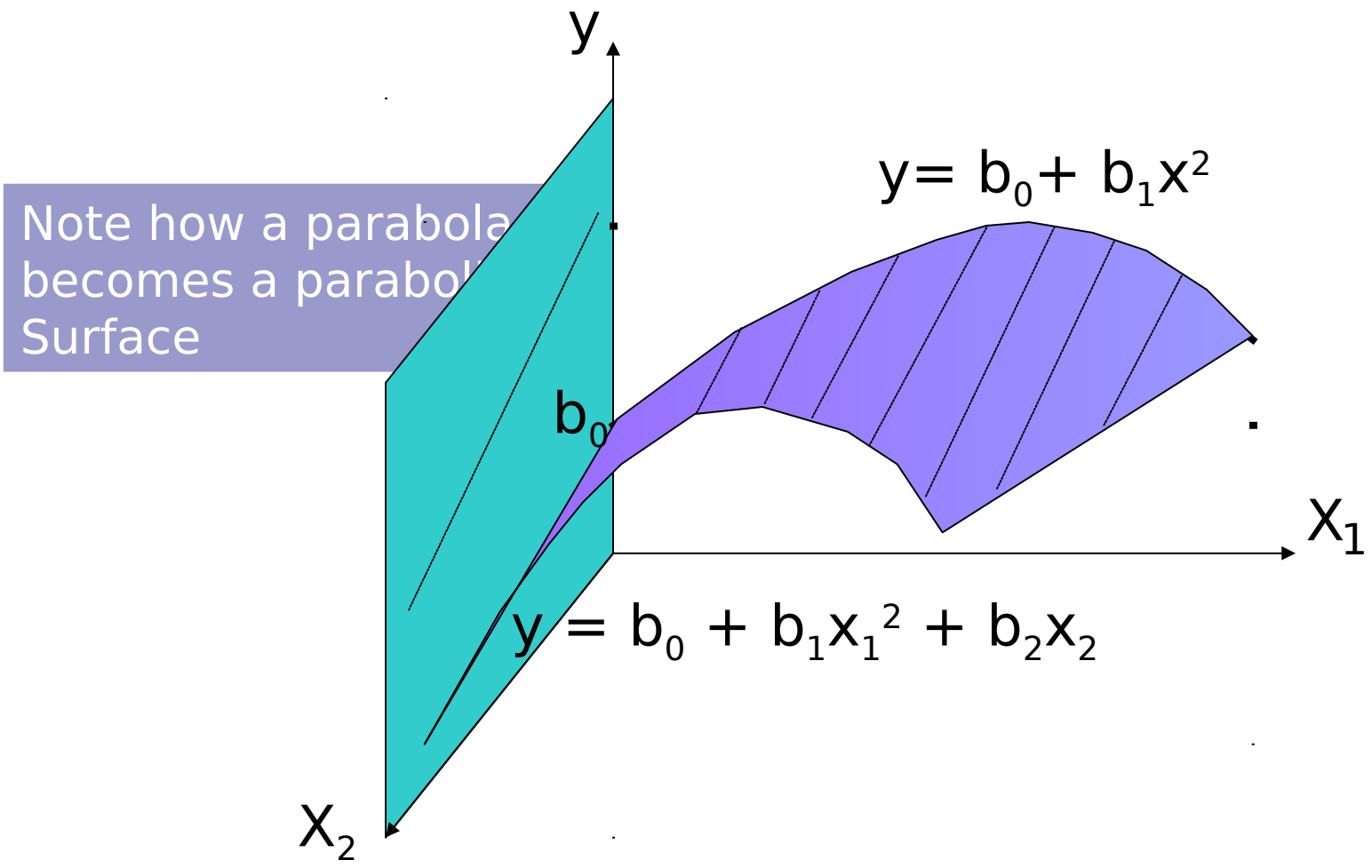
Note how the straight line becomes a plane, and ...



The multiple linear regression model allows for more than one independent variable.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Multiple Regression for k = 2



Logistic Regression

- Regression analysis provides an equation allowing you to **predict** the score on a variable, given the score on other variable(s) assuming adequate sample of participants have been tested
- Linear, Multiple, Logistic, Multinominal
- Example
  - College admissions
  - The admissions officer wants to predict which students will be most successful
  - She wants to predict success in college (*i.e.,* graduation) based on . . .

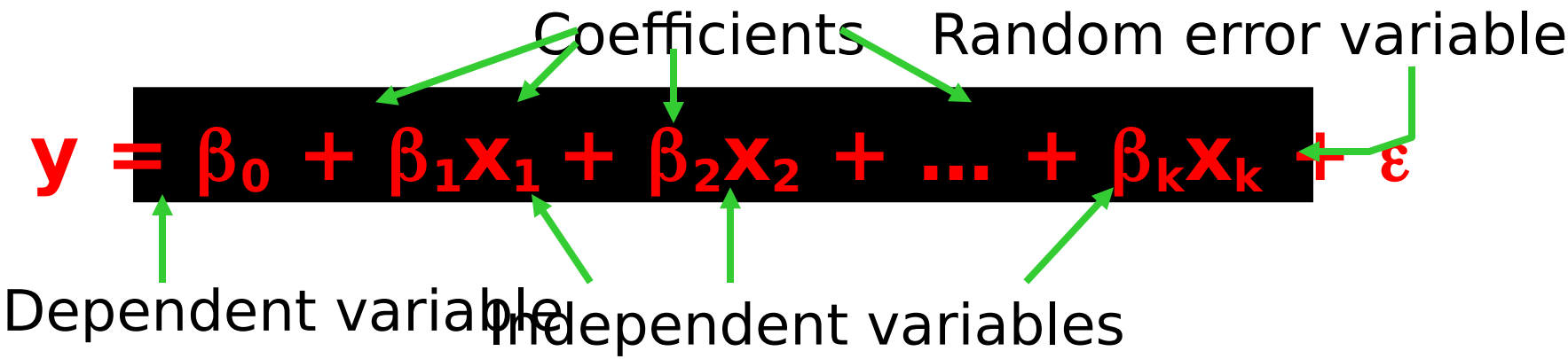
College Success

- GPA
- SAT/CAT
- Letter/Statement
- Recommendation
- Research
- Extra Curriculars
- Luck
- Picture

Model and Required Conditions

We allow for k independent variables to potentially be related to the dependent variable:

$$E(y|x_1,\ldots,x_k)=\beta_1x_1+\ldots+\beta_kx_k$$



College Success

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon$$

where: x_1=GPA, x_2=SAT, x_3=Letters, x_k=Good Looks, e=Luck

$$\mathbf{y = \beta_0 + \beta_1 GPA + \beta_2 SAT + \beta_3 Letters + \ldots + \beta_k Looks + Luck}$$

where: GPA=3.85, SAT=1250, Letters=7.5,Looks=4,Luck=10

$$\mathbf{y = \beta_0 + \beta_1 3.85 + \beta_2 1250 + \beta_3 7.5 + \ldots + \beta_k 4 + 10}$$

where: $\beta_0 = .10$, $\beta_1 = .36$, $\beta_2 = .05$, $\beta_3 = .08$, $\beta_k = .045$

$$\mathbf{y = .10 + (.36 * 3.85) + (.05 * 1250) + (.08 * 7.5) + (.045 + 4) + 10}$$

y = 80.166 with 75 cut-off

Conclusions

- Correlation
- Regression
- Multiple Regression
- Logistic Regression

Questions

Father of Regression Analysis

Carl F. Gauss (1777-1855)

German mathematician, noted for his wide-ranging contributions to physics, particularly the study of electromagnetism. Born in Braunschweig on April 30, 1777, Gauss studied ancient languages in college, but at the age of 17 he became interested in mathematics and attempted a solution of the classical problem of constructing a regular heptagon, or seven-sided figure, with ruler and compass. He not only succeeded in proving this construction impossible, but went on to give methods of constructing figures with 17, 257, and 65,537 sides. In so doing he proved that the construction, with compass and ruler, of a regular polygon with an odd number of sides was possible only when the number of sides was a prime number of the series 3, 5, 17, 257, and 65,537 or was a multiple of two or more of these numbers. With this discovery he gave up his intention to study languages and turned to mathematics. He studied at the University of Göttingen from 1795 to 1798; for his doctoral thesis he submitted a proof that every algebraic equation has at least one root, or solution. This theorem, which had challenged mathematicians for centuries, is still called “the fundamental theorem of algebra” (see ALGEBRA; EQUATIONS, THEORY OF). His volume on the theory of numbers, Disquisitiones Arithmeticae (Inquiries into Arithmetic, 1801), is a classic work in the field of mathematics.

Gauss next turned his attention to astronomy. A faint planetoid, Ceres, had been discovered in 1801; and because astronomers thought it was a planet, they observed it with great interest until losing sight of it. From the early observations Gauss calculated its exact position, so that it was easily rediscovered. He also worked out a new method for calculating the orbits of heavenly bodies. In 1807 Gauss was appointed professor of mathematics and director of the observatory at Göttingen, holding both positions until his death there on February 23, 1855.

Although Gauss made valuable contributions to both theoretical and practical astronomy, his principal work was in mathematics and mathematical physics. In theory of numbers, he developed the important prime-number theorem (see E). He was the first to develop a non-Euclidean geometry (see GEOMETRY), but Gauss failed to publish these important findings because he wished to avoid publicity. In probability theory, he developed the important method of least squares and the fundamental laws of probability distribution, (see PROBABILITY; STATISTICS). The normal probability graph is still called the Gaussian curve. He made geodetic surveys, and applied mathematics to geodesy (see GEOPHYSICS). With the German physicist Wilhelm Eduard Weber, Gauss did extensive research on magnetism. His applications of mathematics to both magnetism and electricity are among his most important works; the unit of intensity of magnetic fields is today called the gauss. He also carried out research in optics, particularly in systems of lenses. Scarcely a

Regression

- As well as describing the type of correlation that may exist between two variables, it is also possible to find the regression line for that scatter diagram (line of best fit)
- When you have two variables it is usual to assign on to be the explanatory variable (independent, x values) - the variable that you have some control over - and one to be the response variable (dependent, y values) - the one you measure that changes because of the explanatory variable
- When calculating a line of best fit in this way, you will work out $y = a + bx$ where y is the predicted value for a give x value (this is regressing y on x)